Anomaly and Signature Filtering Improve Classifier Performance For Detection Of Suspicious Access To EHRs

Jihoon Kim, MS[1], Janice M Grillo, RN, MSIS[2], Aziz A Boxwala, MD[1], Xiaoqian Jiang, PhD[1], Rose B Mandelbaum, MS[2], Bhakti A Patel, BS[2], Debra Mikels, OTRL[2], Staal A Vinterbo, PhD[1], PhD[1], Lucila Ohno-Machado, MD, PhD[1]

[1]Division of Biomedical Informatics, University of California San Diego, La Jolla, CA; [2]Partners Healthcare System, Boston, MA

ABSTRACT

Our objective is to facilitate semi-automated detection of suspicious access to EHRs. Previously we have shown that a machine learning method can play a role in identifying potentially inappropriate access to EHRs. However, the problem of sampling informative instances to build a classifier still remained. We developed an integrated filtering method leveraging both anomaly detection based on symbolic clustering and signature detection, a rule-based technique. We applied the integrated filtering to 25.5 million access records in an intervention arm, and compared this with 8.6 million access records in a control arm where no filtering was applied. On the training set with cross-validation, the AUC was 0.960 in the control arm and 0.998 in the intervention arm. The difference in false negative rates on the independent test set was significant, $P=1.6\times10^{-6}$. Our study suggests that utilization of integrated filtering strategies to facilitate the construction of classifiers can be helpful.

INTRODUCTION

Access to Electronic Health Records (EHRs) by authorized users is now available from virtually anywhere and allows clinicians to make informed decisions that improve overall health care quality. While this availability is great, at the same time it can have associated risks. One potential risk is that these data may be viewed by authorized users for reasons other than providing treatment, payment, or health care operations (*i.e.,* insider abuse or threat is possible). Since the very first IT survey on cyber-attacks[1], one fact has remained almost constant: a greater percentage of attacks comes from the inside (from “trusted” individuals) – 60 to 70 percent – than from the outside (the “untrusted” individuals). Or, to put it another way, roughly twice the number of attacks comes from inside *vs.* outside. Users may abuse their privileges for personal reasons, inappropriately viewing records of relatives, friends, neighbors, co-workers, and celebrities. Most EHR systems on the market today take the form of role-based access gateways where, once the individuals connected have been authenticated, they are allowed complete access inside the perimeter, where they are then free to roam[1].

The participating organizations in this study have several systems in place to provide only the minimum necessary level of access. However, without the ability to understand real-time user/patient relationships, the default procedure for these organizations has been to err on the side of caution and allow the user access to occur. While audit logs provide information to help privacy officers determine if a breach has occurred, there are some fundamental limitations to their practical usefulness in finding potentially suspicious accesses. First, the sheer volume of audit records, which are extracted for manual review by querying the system by user or by patient, make logs more useful as adjuncts to help investigate suspected breaches, rather than tools that can help find inappropriate accesses pro-actively. Second, the audit records themselves do not provide specific situation or relationship information, making it difficult for privacy officers to determine a reason for a particular access. Finally, the manual audit review process is labor-intensive: without knowing where to look for potential breaches, it is hard to find rare cases of inappropriate accesses that are mixed with all of the appropriate accesses.

Several approaches exist to detect inappropriate access to EHRs: (1) Restricting access control[2], (2) Applying patient-user matching algorithms[3], (3) Applying scenario-based rule extraction[4], and (4) Information gathering from EHR and non-EHR systems using a secure protocol[5]. However, none of these approaches have implemented machine learning techniques in a real clinical system. A comprehensive review can be found in our initial study called MAP1 (Monitoring Access Pattern phase 1)[6], where we have shown that statistical and machine learning methods could play a role in the identification of potentially inappropriate access to EHRs. In MAP1[6] we constructed an event data mart of EHR access events from the operational databases collected over a two-month period. Then we defined 26 features likely to be useful in detecting suspicious accesses based on historic breaches. Next we created a training set iteratively in collaboration with privacy officers. Selected events were labeled either as *suspicious* or *appropriate* by privacy officers, and served as the gold standard set for model evaluation. We trained logistic regression (LR) and support vector machine (SVM) models on 10-fold cross-validation sets of 1,291 labeled events. We validated the final model on the external set of 78 inappropriate events that were independently identified by privacy officers. The area under the receiver operating characteristic curve (AUC) of the final model on the whole data set of 1,291 events was 0.911 for LR, and 0.949 for SVM. On the external set, all of which were determined to be *inappropriate*, the sensitivity was 0.759 for LR and 0.793 for SVM at the prediction probability cutoff of 0.5. The results suggested that data mining methods could play an important role in detecting suspicious accesses to EHRs. We report here the extension of this work, call it MAP2, related to fine-tuning of the detection algorithm.

In MAP2, we focus on the construction of classifiers with appropriate filtering technique that can detect rare events, a topic that has received much attention in the informatics literature[7-10]. Signature detection is a rule-based algorithm that constructs a set of rules based on historic breaches. It can correctly detect known patterns and it is easily interpretable[11]. However, it cannot detect unseen patterns and cannot assign risk scores[11, 12]. Anomaly detection compares incoming instances to previously built profiles. It can detect novel patterns, but it requires a large number of historic data. It is relatively difficult to interpret, it cannot assign prediction scores, and it is known to produce too many false positives[13, 14]. Classifier detection relates to determining a classification function based on a labeled training set[15]. It can be fast, accurate, and can assign risk scores to all events[16]. Its downside is that getting the class label is expensive. The ability to score unlabeled events is important in a large-scale data mining, as human validation is limited and costly[17]. Often, reviewers want to sort instances by the risk score and then conduct investigation starting from the highest risk to achieve maximum efficiency in detection. Our approach is to use three-pronged method integrating all three approaches. We extend our previous MAP1 classifier by incorporating an integrated filtering method using anomaly and signature detection.

METHODS

The overview of the study design is depicted in Figure 1. We built a Data Mart of EHR access events from operational databases of two institutions between June 21 and November 29, 2009. Then two population sets were created for control and intervention arms. We applied the MAP1 algorithm[6] to the control arm, and applied our proposed MAP2 filtering technique to the intervention arm. During the study period our privacy officers continued their jobs in parallel to detect suspicious access to EHRs, and they found 78 positive events. We call this the <Parallel Set>, and we used it as a hold-out set for comparing the performance of models induced in the two arms. Figure 1 shows a high level overview of this study, and in-depth explanation is provided in Figure 2 and Tables 1. Institutional Review Board approval was obtained at the Partners Healthcare System.

Control Arm

In the control arm, we applied the MAP1 logistic regression model[6] to obtain prediction probabilities for all events in the population set. Because the events of interest were extremely rare ones, we performed stratified sampling to select some for labeling. Equal number of events at the top (> 0.7), middle (between 0.3 and 0.7) and bottom (< 0.3) tiers were randomly selected and were sent to privacy officers. An event was labeled as positive if it was highly suspicious and required thorough investigation and the examination of post-hoc evidence in order for privacy officers to determine whether or not it constituted a breach. We obtained a labeled set of the record access event

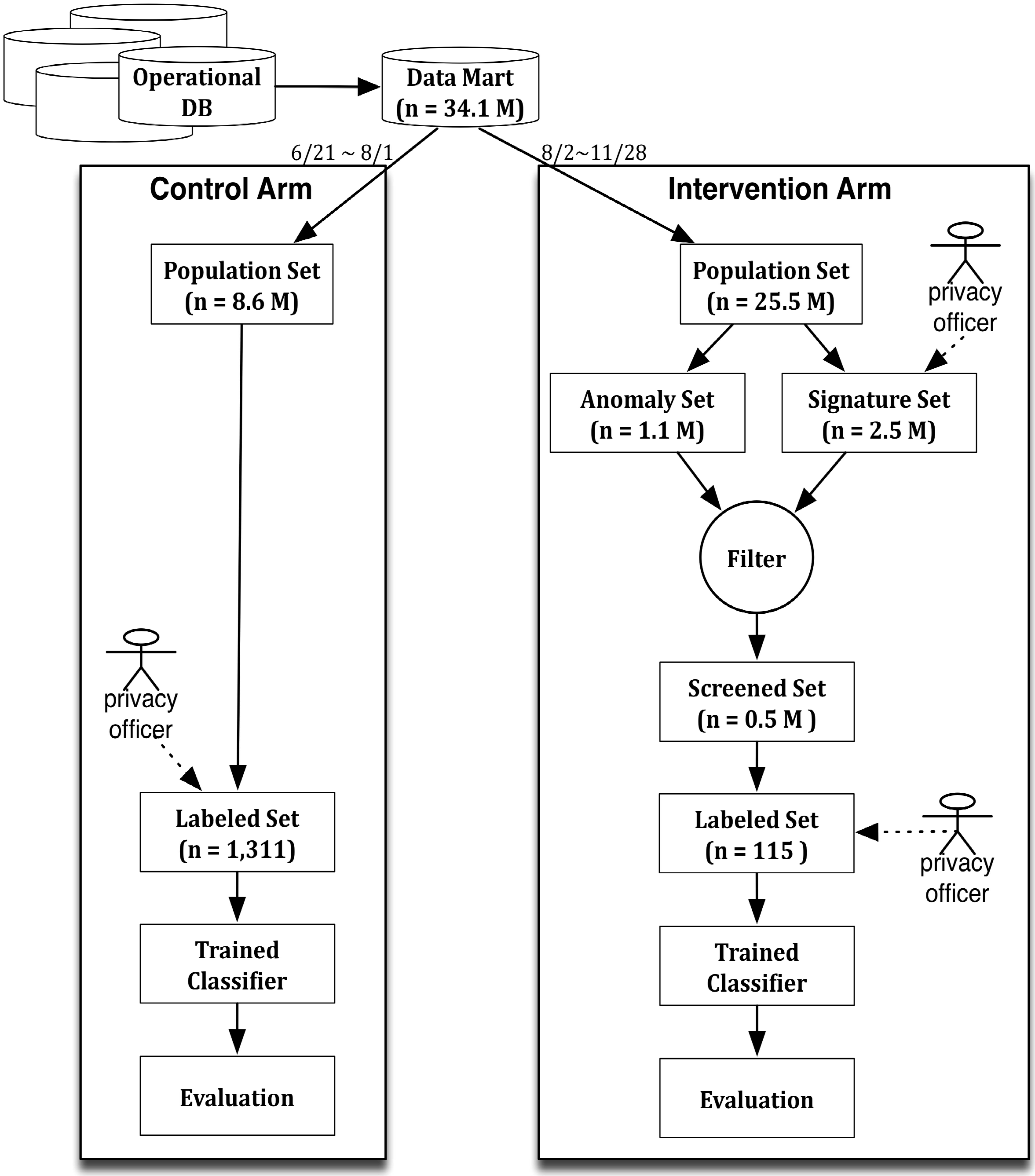

Figure 1. Overview of Study Design. The labeled set in the control arm was built using the accesses to EHRs occurred during 6 weeks between June 21 and August 1 of 2009. While the labeled set in the intervention arm was constructed with the accesses occurred in 4 weeks between November 1 and 28, the historic data of the recent 90 days (August 2 ~ October 31) were used for user-profiling and department-profiling in the anomaly set. In the intervention Arm, additional three data sets, <Anomaly Set>, <Signature Set> and <Screened Set>, are used for filtering. None of these appear in the Control Arm, where no filtering was applied. Sample events were sent to privacy officers who provided labels (show in dotted line). Privacy officers also supported the data mining team to extract signature types based on previous breaches.

this way. Next a logistic regression model was trained and was evaluated on the hold-out <Parallel Set>.

Intervention Arm

Signature Set

We reviewed past privacy breaches and extracted six rules (or 'snooping types'). An event was tagged with snooping type names whenever it met the following conditions at the time of access:

- COW (for coworker) if both the user and the patient worked in the same department
- EMP (for employee) if the patient was an employee of the institution
- FAM (for family) if the user and individuals listed by the patient as contacts had the same last names
- LDE (for long-deceased patient) if the patient had died more than a year before the time of access
- NEI (for neighbor) if the user and the patient had the same zip code
- VIP if the patient is VIP flagged patient in the system.

Each rule was implemented as an SQL-query with multiple conditions and at least one snooping type was present in 2.5 million events. We called this the <Signature Set>.

Anomaly Set

We defined a BasePattern to be a covariate pattern[18] value of four key features {Care unit visit match, clinic match, MRN match, patient had recent visit}. As each of four features was binary, a BasePattern could have 16 different possible values {0, 1,…, 15} after binary-to-decimal conversion. We assigned a BasePattern for each event in the population set of the intervention arm. Once a BasePattern was assigned to each access event, symbolic clustering of access events was completed. Then, for any given event, we counted the frequency of base patterns per user for all user events occurred during the most recent 90 days. We called this process "user-profiling". An event was marked as anomaly at the user-level if the observed proportion of its base pattern was less than 5% among all the user events occurred during the past 90 days. Similarly, we counted the frequency of base patterns for all events within a department in during the most recent 90 days. We called this process "department-profiling". An event was marked as anomaly at the department-level if the observed proportion of its base pattern was less than 5% among all the department events occurred during the recent 90 days. An event was marked as anomaly if it was abnormal either at the user-level or the department-level. There were 1.1 million events. We call this <Anomaly Set>.

Screened Set

While most negative accesses were alike, each negative access was on its own, according to our previous study[6], which confirmed our hypotheses that most staff individuals accessed EHRs for job-related reasons. Also, from a practical stand-point, our privacy officers wanted to spend their time on high-risk accesses. This class-imbalanced problem and its solution have been addressed by machine learning investigators[19, 20]. We used BasePattern, <Rare Set> and <Screened Set> to derive four suspicion levels of access events in four color-codes, with the suspicion order of Red > Orange > Yellow > Green, as shown in Table 1. We chose the highest risk categories, red and orange events, as our <Screened Set>.

Mixed Set

It would be ideal if we could directly compare classifier performance between two arms where each labeled set had positive and negative access records. However, all sampled events in the intervention arm turned out to be positive ones, making evaluation on the labeled set alone impossible. We solved this problem by creating a mixed set: "borrowing" negatives accesses from the control arm (Figure 2). We copied all the negative access of the labeled set in the control arm to the mixed set. Then we sampled the positives of the labeled set from the intervention arm and added to the mixed set. The sample size of positive accesses was equated to that of positive access in the control arm. This ensured controlling for confounding factor of the sample size to the classifier performance. We repeated the construction of the mixed set 30 times.

Suspicion		Defining Features		Selected for
Level	Has BasePattern less than 4	Has any signature type	Has any anomaly pattern	Screened Set
Red	Yes	Yes	Yes	4
Orange	Yes	Yes	No	4
	Yes	No	Yes	4
Yellow	Yes	No	No	
	No	Yes	Yes	
	No	Yes	No	
	No	No	Yes	
Green	No	No	No	

Table 1. Construction of the Screened Set. Three binary (yes/no) features were derived and grouped in suspicion levels using a four-colored code. Then two high-risk groups were selected to form the Screened Set.

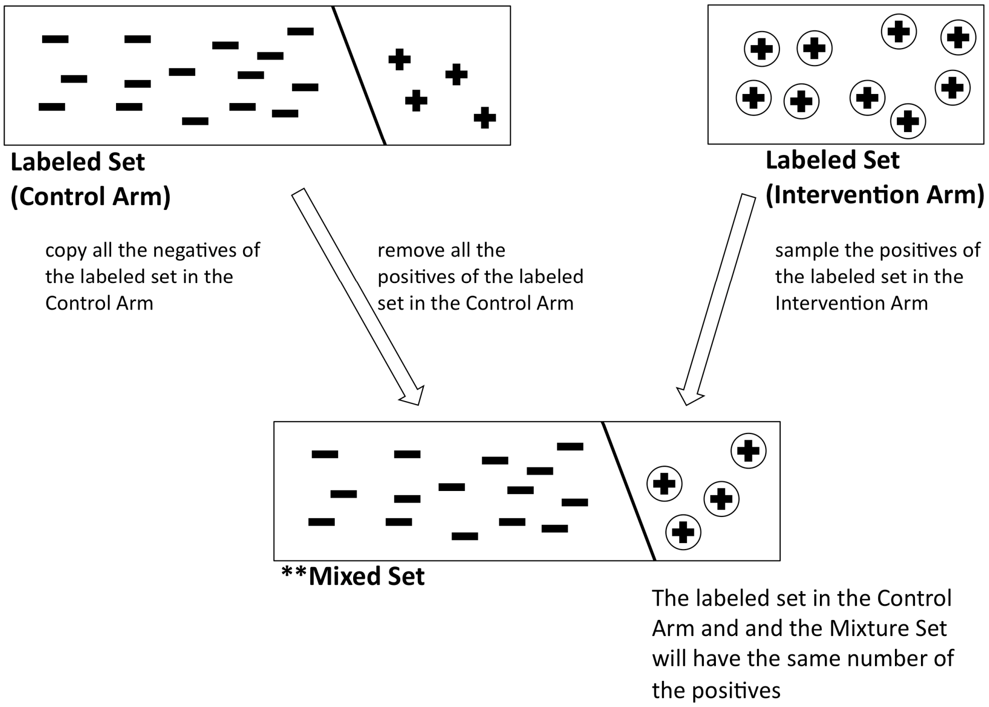
Labeled Set
(Control Arm)
Labeled Set
(Intervention Arm)
copy all the negatives of the labeled set in the Control Arm
remove all the positives of the labeled set in the Control Arm
sample the positives of the labeled set in the Intervention Arm
**Mixed Set
The labeled set in the Control Arm and and the Mixture Set will have the same number of the positives

Figure 2. Construction of Mixed Set. As the sampled accesses in the intervention arm all turned out to be positives, a classifier could not be trained. Hence, a new data set was constructed by mixing the positives inherited from the intervention arm and the negatives "borrowed" from the control arm. The numbers of negatives and the positives in the mixed set were equated with those in the control arm to take into account the sample size effect

Evaluation

To compare classifier performance between two arms, we used the Area Under the ROC Curve (AUC) and false negative rates on the labeled set of the control arm and on the 30 mixed sets. An ideal evaluation would be the one with a prospective dataset, where unlabeled events with a corresponding prediction are further labeled. However, due to the resource limitations, we could not perform it. The hold-out test set had all positive class labels, as it was derived from documented inappropriate access to records as determined by privacy officers, so we had to use false negative rate as an evaluation measure. Also, visual diagnosis was conducted on the distribution of prediction probabilities per class (positive or negative) on both training set and the mixed set.

RESULTS

Table 2 displays breakdown of number of accesses per data-set. The control arm had 8,580,190 accesses collected over six weeks between June 21 and August 1, 2009. The intervention arm had 25,504,815 accesses over 17 weeks between August 2 and November 28, 2009. Actual sampling and labeling in the intervention arm was conducted during November 2009. Thirteen weeks' data prior to November were used to conduct user-profiling and department- profiling using the events occurred during recent 90 days. While the labeled set in the control arm had both positives and negatives (59 vs. 1,252 respectively), the labeled set in the intervention arm had only positives (n=115). During study period (June 21 ~ November 28, 2009), privacy officers conducted detection of inappropriate accesses in parallel. They found 78 events, all positive, which were not revealed to the data mining team. We held out 78 <Parallel Set> events from the population sets and used as a test set for evaluation.

Table 3 shows the number of accesses affected by anomaly and signature filtering in lattice. At the user level, 21,762 distinct users had user-lever cluster size between 1 and 11 and the most frequently observed cluster size was 3. At the department level, 1,597 distinct departments had cluster sizes between 1 and 12 and the most frequently observed was 5. <Anomaly Set> had 1,131,349 accesses. Among six signature types, VIP had the largest number of accesses, and COW had the smallest.

Name	Label	Control Arm	Intervention Arm	Dates 6/21 ~ 8/1	Dates 8/2 ~ 10/31	Dates 11/1~ 11/28	Number of Accesses
Control Arm							
Population Set	No	○		4			8,580,190
Labeled Set	Yes	○		4			1,311
Intervention Arm							
Population Set	No		○		4	4	25,504,815
Anomaly Set	No		○		4	4	1,131,349
Signature Set	No		○			4	2,578,782
Screened Set	No		○			4	475,923
Labeled Set	Yes		○			4	115
Parallel Set	Yes			4	4	4	78

Table 2. Breakdown of number of accesses. The <Parallel Set> is a hold-out set for evaluation that is not used in the control arm nor the intervention arm.

Signature Type	**Anomaly Pattern**			**Common Pattern**	**Total***
	User	Department	Any (User or Dept.)		
Co-worker	285	165	326	2338	2,664
Employee	23,902	29,168	35,182	809,303	844,485
Family	5,298	4,596	6,862	26,732	33,593
Long-deceased	7,821	14,490	15,683	30,214	45,897
Neighbor	3,381	3,709	4,393	60,658	65,051
VIP	54,576	68,366	78,071	1,513,117	1,591,188
No snooping type	616,484	857,354	992,196	21,933,837	22,926,033
Total**	711,017	976,606	1,131,349	24,373,466	25,504,815

Table 3. Number of events per signature type and anomaly/common pattern. *The row totals do not add up to the total, as each access can have more than one snooping type. **Two sub-totals, 'any anomaly pattern' and 'common pattern' add up to the final total.

Evaluation

The AUC on the control arm was 0.960 and the mean AUC of 30 mixed sets in the intervention arm was 0.998. The difference in AUC between the two arms was significant ($P = 1.8 \times 10^{-6}$). The false negative rate on the <Parallel Set> was 0.508 on the control arm, and the mean false negative rate was 0.062 on the 30 mixed sets. The difference in the false negative rates on the <Parallel Set> was also significant ($P = 1.6 \times 10^{-6}$). The false negative rate in the control arm had only one value and it is drawn as a dotted line. On the other hand, the intervention arm had 30 false negative rates, hence a box plot was drawn with its median value on the bold horizontal line.

We also conducted visual diagnosis of class distribution of prediction probabilities to see if the classifier trained on the control arm was usable. Figure 3 displays a series of probability distributions on the training set and the test set per class. Here, the training set is the labeled set of the control arm, and the test set is the first mixture set of the intervention arm. The first mixed set is representative of all 30 based on the their small standard deviations of AUC and false negative rates. A perfect classifier would have produced all the negatives as 0 and all the positives as 1. According to the top histograms in Figure 3, all the negative show near-perfect predictions in both control and intervention arms. However, histograms of negatives in the control arm are spread all over the range 0 and 1. At the bottom, even though all events of test set, or <Parallel Set>, are confirmed positives, 35.9% of them have prediction probabilities lower than 0.5. The trained classifier on the control arm would have missed their detection. On the other hand, the distribution of the positives in the intervention arm shows distinct skewness toward 1. In the intervention arm, 11.5% of events had prediction probabilities lower than 0.5 on the test set, and 10.2% on the training set.

Case study

We consulted privacy officers to find out why our intervention arm approach missed those three events out of 78 in the <Parallel Set> using a cutoff probability 0.5. The first event missed in the intervention arm had a probability estimate 0.001. We missed this event because this patient was not marked in our database as an employee, even though the privacy officer reported the incident as a co-worker access breach. In addition, this was not a rare pattern for the user. We do not know why the patient was not marked as an employee. It is possible that this user was a consultant, or may have been hired shortly after this event took place. There was also a clinic match associated with this access, which is most often associated with appropriate accesses. The second and the third missed events in the intervention arm had the same probabilities: 0.074. They belonged to the same user-patient pair. In this case, the patient filed a complaint that her family name match for these two accesses, but no address or zip match. In addition, the patient had recently been seen at the clinic where her mother worked, making this access fit the typically appropriate, and very common, pattern of clinic match access, but no address of zip match.

Impact

A major difference between MAP1 and MAP2 was that, for MAP2, investigations were conducted and sanctions were imposed where necessary. Based on MAP2 results, an employee education campaign was conducted reinforcing the Minimum Necessary policy, specifically reminding staff that family member accesses were not appropriate. Based on investigations, users were sanctioned where they were found to have violated the policy. This did include termination in one case. Other offenders were given warnings and were re-educated about the policies. We proposed a framework for rare event detection in disparate hospital systems using large-scale data. The same method could be applied to rare event detection problems in other domains, such as detecting high-risk patients likely to have rare complications after surgery, adverse drug events, cases of medical insurance fraud, etc.

Limitation

We have tested our system in two tertiary care systems in New England. It is possible that some of the pattern we described may not all be found in other medical centers, and that other patterns may exist that we were not able to capture in our study. Additionally, the system is not fully automated, and the utilization of privacy officer was still necessary to label cases in high-risk categories. Given how time-consuming the process of labeling one event is (multiple phone calls, emails, interview with the suspect's supervisor, examination of records in multiple disparate systems, etc), the number of labeled accesses is still small.

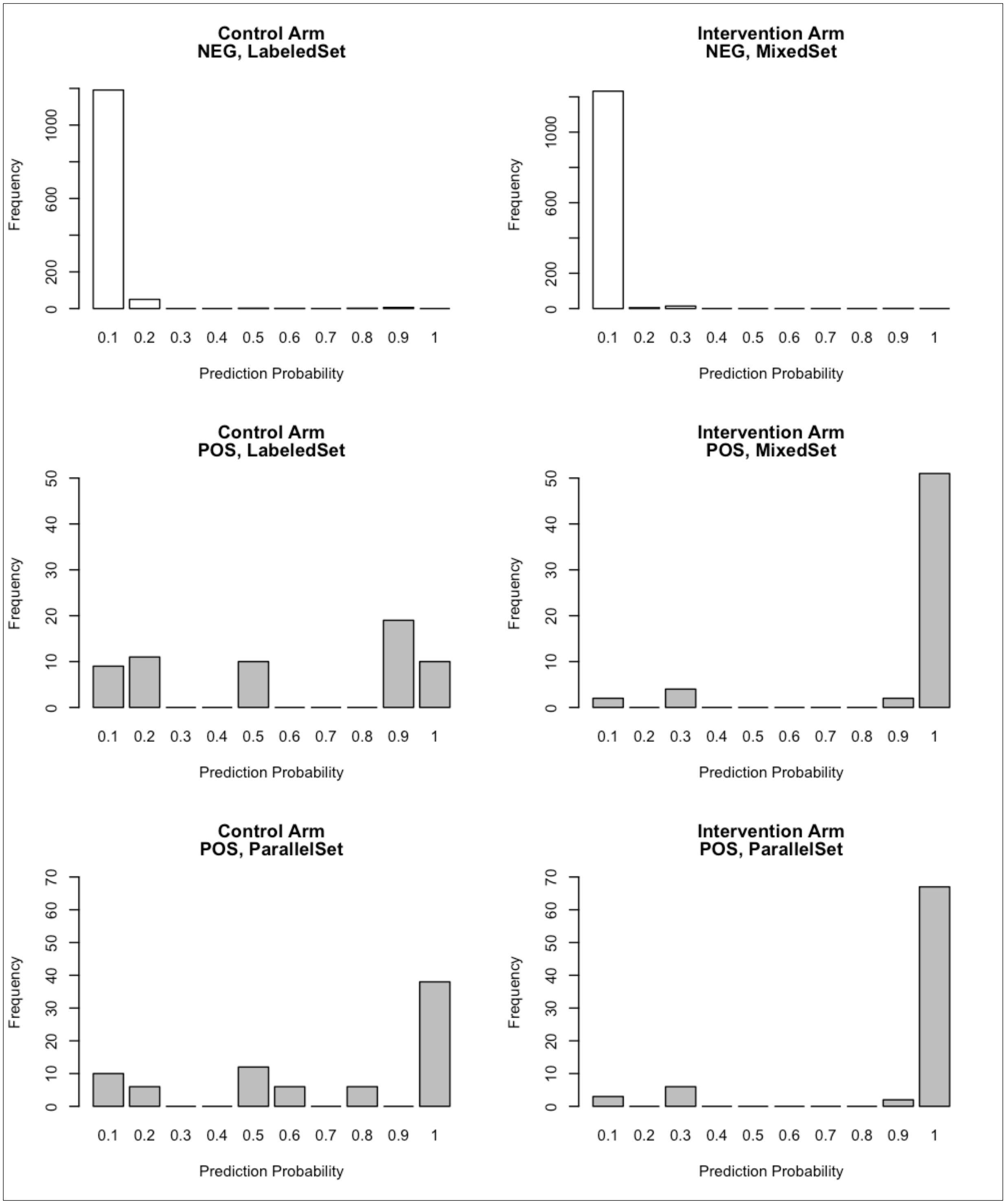

Figure 3. Distribution of prediction probabilities. The first column represents the control arm and the second column the intervention arm. The top rows represent distributions of prediction probabilities for negative events (=NEG, white bars). The middle and bottom rows represent distributions of prediction probabilities for positive events (=POS, shaded bars).

CONCLUSION

We described an integrative filtering method leveraging anomaly detection, signature detection and a rule-based technique that could assist with detection of suspicious accesses to EHRs. In cases in which labeling of events is extremely time consuming, positive cases are extremely rare, and the volume of data is overwhelming, the approach described here can be used to facilitate the detection of rare, but important, events.

Acknowledgements

This study was funded in part by the Partners Information Systems, the Partners Siemens Research Council and R01LM009520, NLM, NIH. We acknowledge K Grant, P Rubalcaba, C Spurr, D Adair, J Raymond, J Einbinder, J Crowley-Smith, H Shea, M Hamel, C Griffin, J Saunders, P Moran, M Miga, M Muriph, A Kennedy and L Damiano for their contributions to this research.

REFERENCES

1. Lynch DM. Securing against insider attacks. Information Systems Security. 2006;15(5):39-47.
2. Ferreira A, Cruz-Correia R, Antunes L, Chadwick D. Access control: how can it improve patients' healthcare? Stud Health Technol Inform. 2007;127:65-76.
3. Salazar-Kish J, Tate D, Hall PD, Homa K. Development of CPR security using impact analysis. Proc AMIA Symp. 2000:749-53.
4. Asaro PV, Herting RL, Jr., Roth AC, Barnes MR. Effective audit trails--a taxonomy for determination of information requirements. Proc AMIA Symp. 1999:663-5.
5. Malin B, Airoldi E. Confidentiality preserving audits of electronic medical record access. Stud Health Technol Inform. 2007;129(Pt 1):320-4.
6. Boxwala AA, Kim J, Grillo JM, Ohno-Machado L. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. J Am Med Inform Assoc. 2011 Jul 1;18(4):498-505.
7. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med. 2006 May;37(1):7-18.
8. Ohno-Machado L. Identification of low frequency patterns in backpropagation neural networks. Proc Annu Symp Comput Appl Med Care. 1994:853-9.
9. Ohno-Machado L, Musen M. Learning rare categories in backpropagation. Advances in Artificial Intelligence. 1995:201-9.
10. Owen AB. Infinitely imbalanced logistic regression. The Journal of Machine Learning Research. 2007;8:761-73.
11. Barbara D, Jajodia S. Applications of data mining in computer security: Springer Netherlands; 2002.
12. Lee W, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. 1999: IEEE; 1999. p. 120-32.
13. Das K, Schneider J, Neill DB. Anomaly pattern detection in categorical datasets. 2008: ACM; 2008. p. 169-76.
14. Patcha A, Park JM. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks. 2007;51(12):3448-70.
15. Shen A, Tong R, Deng Y. Application of classification models on credit card fraud detection. International Conference on Service Systems and Service Management; 2007: IEEE; 2007. p. 1-4.
16. Mukkamala S, Janoski G, Sung A. Intrusion detection using neural networks and support vector machines. 2002: IEEE; 2002. p. 1702-7.
17. Zhu J, Wang H, Yao T, Tsou BK. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. 2008: Association for Computational Linguistics; 2008. p. 1137-44.
18. Hosmer DW, Lemeshow S. Applied logistic regression: Wiley-Interscience; 2000.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16(1):321-57.
20. Japkowicz N. The class imbalance problem: Significance and strategies. 2000: Citeseer; 2000. p. 111-7.